

White Paper Report

Report ID: 102663

Application Number: HD5130111

Project Director: Benjamin Vershbow (benjamin_vershbow@nypl.org)

Institution: New York Public Library

Reporting Period: 4/1/2011-12/31/2012

Report Due: 3/31/2013

Date Submitted: 4/1/2013

**National Endowment for the Humanities
Final White Paper Report
Grant HD-51301-11**

**Crowdsourcing Culinary History at The New York Public Library
April 1, 2011—December 31, 2012**

**Project Director: Benjamin Vershbow
The New York Public Library
March 31, 2013**

Final Report

Background

In April 2011, the National Endowment for the Humanities (NEH) awarded a \$50,000 grant to The New York Public Library (NYPL) to build an online platform that would build the Library's capacity to crowdsource transcription of its unique collection of historic restaurant menus. NYPL is pleased to report that the project has been successfully completed, and this final report outlines the Library's work on the project from April 1, 2011 through December 31, 2012.

Overview

*What's On The Menu?*¹ was designed to transform roughly 9,000 digitized images from NYPL's famed Buttolph Menu Collection into a searchable database of historical culinary and economic trends. Because of difficulties in mechanically extracting quality text from the menus, and from a desire to build a *structured* data set of discrete dishes and prices, it was determined that manual transcription would be the best method for creating the database. As this task was far too large and time-consuming to accomplish with internal resources, the Library saw this situation as an opportunity to explore user collaboration, or 'crowdsourcing', as a means of accomplishing the work. In other words: build a simple web application for transcribing the menus and see if members of the public would be willing to volunteer their time.

The results were astonishing. In the first week after the project's launch, 12,000 unique visitors transcribed over 65,000 dishes from approximately 800 menus. Nearly two years later, nearly 1.3 million dishes were transcribed from over 16,700 menus. Members of the public have spent over 10,583 hours transcribing dishes from menus - or over 1.2 years of cumulative time.

Our goal for this grant was to fund the development of the *What's On The Menu?* web application by an outside software developer and to fund project management in its first six months of public life. However, several fortuitous events occurred which changed the shape of the grant, but not the project. First, the project directors made the decision to build the application with in-house technology staff, netting a large cost savings and ensuring compatibility and institutional knowledge moving forward. Second, the project launched to an overwhelmingly positive response, achieving the very optimistic project goals (9,000 menus transcribed in 6 months) in the first 3 months. Finally we were quite fortunate that the NEH enabled us to modify and extend the grant, which gave us the flexibility to respond to the overwhelming rate of transcription by a newfound community of foodie historians by taking on critical community management tasks and producing the requisite metadata so NYPL could digitize more menus for our transcribers.

Bigger Picture: So You've Digitized a Collection...Now What?

In recent decades, libraries, archives, and museums have been digitizing large quantities of rare and unique materials, moving many of them out of the institution and onto the Internet. While these efforts have greatly expanded basic access to certain collections, additional work is often required to fully realize their usefulness in the digital environment. Where descriptive metadata exists, it often requires modification and normalization against existing standards in order to better surface the materials in federated searches and other data-driven discovery mechanisms.

In some cases, metadata must be created for the first time. In other instances, as with the Menus project, materials must be intensively processed following digitization in order to make their richest contents accessible to web browsers, search engines, and application developers. Historical maps must

¹ <http://menus.nypl.org>

be “georectified” in order to be searched and queried like today’s digital maps². Books and newspapers must be optically processed to extract searchable text, the output of which process often requires subsequent manual cleanup in order to become useful³. Still other texts (such as old city and business directories, church registries, playbills, ship’s manifests, etc.) are in essence databases in printed form, and their contents must be laboriously transformed into structured information in order to be accessed and referenced against other data sets. It’s not enough to simply have that metadata, but finding ways to make it actionable, so they don’t remain simply static images, is a priority.

Migrating the sum of human knowledge to the Internet is daunting, especially for resource-strapped cultural organizations. But these institutions, while lacking the financial and engineering assets of big technology and media companies, have one ace up their sleeve: their public mission. The web contains harbingers of new kinds of public libraries and museums, such as the Internet Archive, Project Gutenberg, Freebase, Wikipedia, Flickr Commons, and the Creative Commons; their purpose is to provide knowledge and tools for the civic realm, to serve as free resources for a lifetime of learning.

Most of these initiatives were built collaboratively, often for no pay or material reward, by a motivated corps of well-informed, diligent enthusiasts looking to devote their time to something bigger than themselves. Through networked collaboration on the Internet and an inspiring mission, Wikipedia built an open access encyclopedia by tapping into the same civic impulses and irrepressible curiosity that have propelled people through the doors of libraries and museums for generations. Now traditional institutions must engage this Internet citizenry in new ways—and many already are.

For NYPL, this suggests a new kind of public library, one built not only *for* but in collaboration *with* its public. The buzzword for this is “crowdsourcing”: breaking up large tasks into small pieces for a dispersed pool of workers. Increasingly, and in large part through the lessons learned in *What’s on the Menu?*, we are coming to see crowdsourcing not only as way to accomplish work that might not otherwise be possible, but as an extension of our core mission. As Trevor Owens, Digital Archivist at the Library of Congress, puts it: “*it is about offering your users the opportunity to participate in public memory.*”⁴

Project Genesis

Founded as a repository for predominantly print and paper-based materials, NYPL is now exploring how to convert an analog knowledge base into a digital resource of comparable import. Mass digitization of the Library’s printed book corpus began nearly a decade ago with its involvement in the Google Library Project and the Open Content Alliance (now absorbed by the Internet Archive). Our internal digitization efforts have focused on special collections: prints, photographs, manuscripts, and other rare book and archival materials. This also has been the locus of NYPL’s experiments in digital collection building and data processing, which lately have been spearheaded by a recently-formed group, NYPL Labs⁵.

The Buttolph Menu Collection was identified early on as a promising test bed for experimentation: an ephemeral “edge” collection that has historically been difficult to catalog and preserve, yet which contains, in the aggregate, vast quantities of cultural data, a potential goldmine to historians, journalists, chefs, novelists, and educators. Embedded in ink and print are millions of interrelated data points that tell social histories as vast as the population of oysters in New York City and as particular

² See Old Maps Online directory of georeferencing tools - <http://help.oldmapsonline.org/georeference>

³ See Google reCAPTCHA - <http://www.google.com/recaptcha> ; also Trove Australian Newspapers - <http://trove.nla.gov.au/ndp/del/home>

⁴ Trevor Owens, “Crowdsourcing Cultural Heritage: The Objectives Are Upside Down” - <http://www.trevorowens.org/2012/03/crowdsourcing-cultural-heritage-the-objectives-are-upside-down/>

⁵ <http://www.nypl.org/collections/labs>

as the price of a cup of coffee at Child's Lunch Room on 130 Broadway in 1901 (5 cents). How best to extract this data? The menu collection has the added advantage of popular appeal. Featured in high-profile exhibitions at NYPL over the past decade and a major source for various scholarly works, the collection has been a consistent draw to enthusiasts, educators, and researchers, tapping into broad popular interest in food and the origins of dining culture.

The aim with *What's on the Menu?* (WOTM) was to create an experience that was simple and to capitalize on popular interest in food to elicit greater participation. In order to attract the broadest possible audience, we strove to keep the task basic and clear (capture only dishes and prices); build the most stripped-down tool possible (click on a region of a menu image, type what you see); and keep barriers to participation low (no registration or login required).

Project Activities and History

What's on the Menu? was born out of a discussion about the menu collection and its value as an economic dataset. Several years before, they had seen a researcher meticulously analyze the records of hundreds of menus, recording the price of fish on menus, to use as a proxy for the fluctuations of fish populations over time, and realized that this kind of research could be performed by anyone if the library offered the collection as a structured dataset. Not only did this kind of thinking resonate with the technologists, it got them excited. It became their side-project.

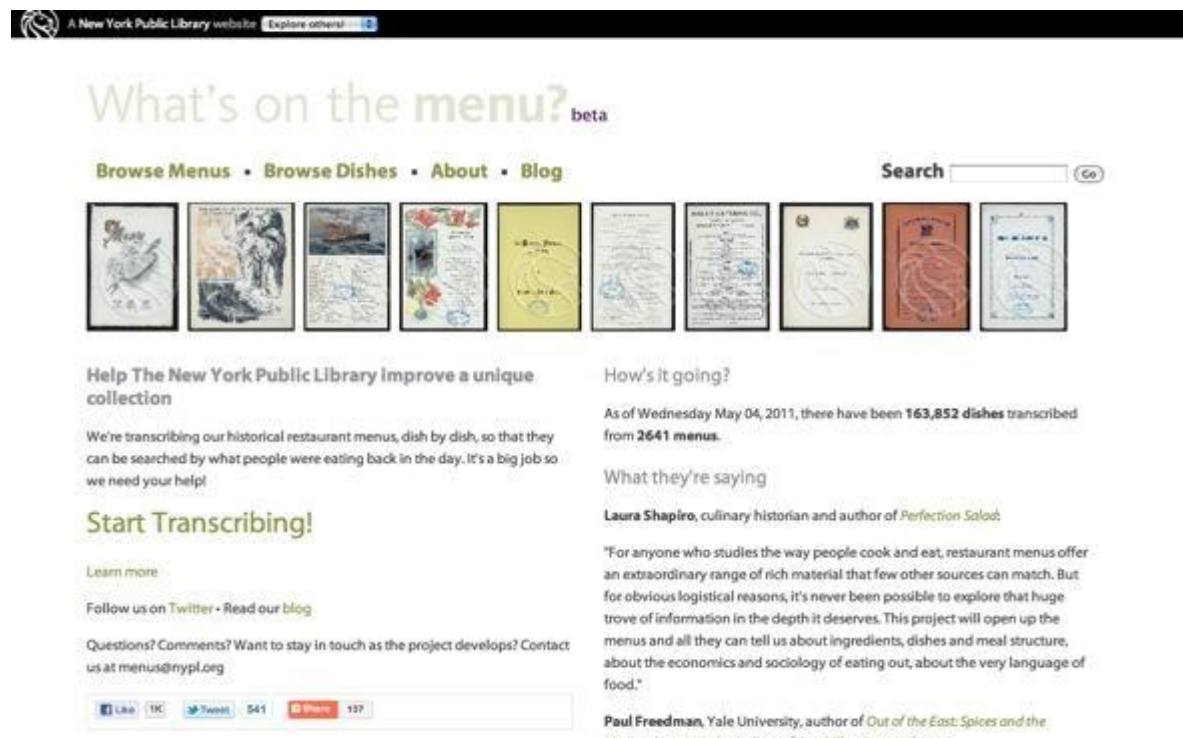
A group of NYPL staff began working in their off-hours to build the site. At the heart of their work was developing a simple interaction for the project's users to perform; a simple action designed to generate a simple output.

1. User sees an image of a historic menu
2. User clicks on a dish
3. User is presented with a zoomed-in image of that selected dish
4. User is asked transcribe exactly the name of the dish, and if present, the price

Many users could participate at once, collaboratively transcribing a single menu together in real-time. This simplicity required a staggering amount of complexity to pull off, and the technology team at NYPL took advantage of some new features of the library's technology stack. The Menus collection was one of the few digitized collections where all assets were stored in the JPEG2000 format and saved in a lossless (uncompressed) form. Because of this, the Library was able to implement a JPEG2000 server, Djatoka, that allowed derivative menu images to be generated on-the-fly. That meant creating high-resolution zoomed areas of the image to transcribe wouldn't need to be generated beforehand and could be created exactly where the user clicked their dish⁶. The core application was built from the ground up as a new program using the Ruby on Rails framework, as we realized early on that even if we used someone else's existing transcription product, our desired interaction was so unique that to implement it smoothly, we'd have to build it ourselves.

A basic first version of the participatory transcription application launched on April 18, 2011 and was an instant hit. In the first ten days, 100,000 dishes were transcribed. The project was featured in *The New York Times* 'Dining Section', National Public Radio, and a number of food and library blogs.

⁶ The continued use of the JPEG200 server on other projects, coupled with the growth of menus has shown that there are significant scalability issues with generating all derivative images on-the-fly as the computational expense of each new image can often exceed the amount of time users are willing to wait to see it.



What's on the Menu? home page, two and a half weeks after launch

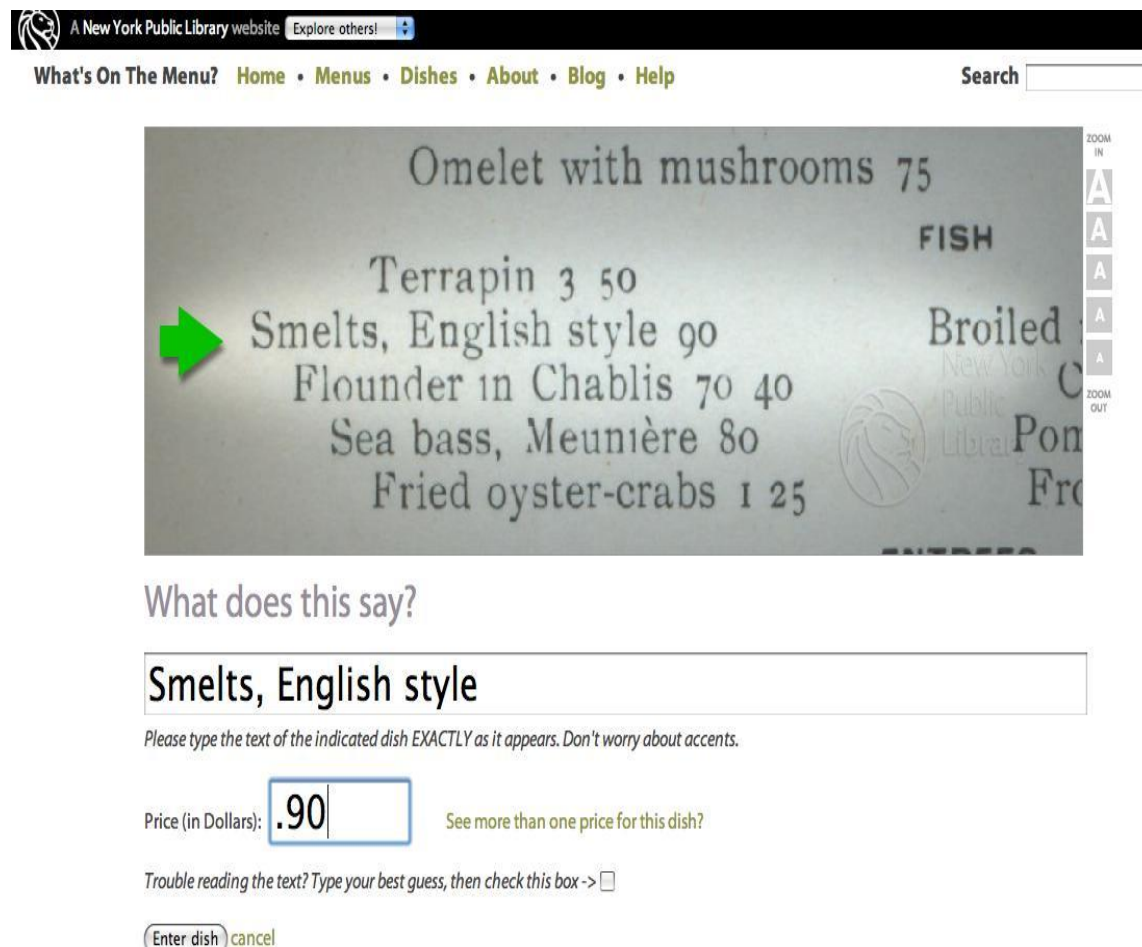
In the following months, New York chefs such as Mario Batali (*Babbo*, *Del Posto*, *Eataly*) and Brooklyn's Doug Crowell (*Buttermilk Channel*) publicly endorsed the project, and others (such as Rich Torrisi of Little Italy's *Torrisi Italian Specialties*, and David Chang, creator of the *Momofuku* chain) created special menus derived from items in the collections. As a testament to the ease-of-use of the tools, a class of 4th graders in San Antonio, Texas used the site to practice typing skills while learning about historical cuisine.

Some of the highest praise came from *The Atlantic's* Technology Editor, Alexis Madrigal, who on June 20, 2011 wrote:

"The original menu has become another node on the living web, you can see "Tutti Frutti" on a menu from 1900 and then make it with a recipe from 1906 or 1962. I think that's brilliant, but not because it's digital. Digital is merely the precondition. The web and search algorithms and crowdsourcing and all that make it possible, but to focus on them would miss the point.

*The point is: this project changes our relationship with time. When we weave history into the web, we weave the past into the present. And that is awesome and important."*⁷

⁷ Alexis Madrigal, *The Atlantic*, "What Big Media Can Learn From The New York Public Library", <http://www.theatlantic.com/technology/archive/2011/06/what-big-media-can-learn-from-the-new-york-public-library/240565/>



What does this say?

Smelts, English style

Please type the text of the indicated dish EXACTLY as it appears. Don't worry about accents.

Price (in Dollars): See more than one price for this dish?

Trouble reading the text? Type your best guess, then check this box -> ☐

Transcription interface

The project was having no trouble keeping participants engaged. In fact, keeping up with demand became the new challenge WOTM had begun with a portion of the collection (a little under a quarter of the total ~45,000 items) that had been digitized some years back for the launch of the NYPL Digital Gallery⁸. With the public making quick work of that initial batch, existing digitization queues had to be reprioritized to ensure a steady flow of new menus.

Effectively re-starting the cataloging of the menus exposed deficiencies in workflows and data models. Rudimentary records (created, in an interesting anticipatory echo of WOTM, by a group of on-site volunteers a decade earlier) already existed for many of the un-digitized menus, but they were inconsistently formatted, and in some cases simply non-existent; even to remediate the existing data would require sorting through each individual menu by hand. A meeting was held with the project directors, the application developer, and the Library's metadata team and it was decided that it ultimately made more sense to retire the legacy menu database and to re-catalog the menus according to a new, lighter-weight schema. This process was put into effect in mid-summer of 2011. We were deep in the complexities of creating a sustainable digital collections project with real-time public demand. The public-facing tools were working well. We were approaching half a million dishes transcribed.

⁸ NYPL Digital Gallery – <http://digitalgallery.nypl.org>

All the while, the project focused heavily on managing a new community of volunteers. As part of the “keep things simple” strategy, the site did not have a user registration system, so all contacts with users were initiated voluntarily through email and social media. We quickly set about recruiting a small team of interns in Fall 2011, who were trained to manage the day-to-day content administration duties: moving menus to the public transcription queue, checking contributions in the site’s “under review” section (the intermediate stage in the menu transcription workflow), and responding to user queries via email.

We began to develop some basic active engagement strategies. Social media was and continues to be our main outreach tool, particularly Twitter and Facebook (as well as good old-fashioned email). These channels help sustain public interest by providing staff with an outlet for sharing discoveries and curiosities from the archive (peculiar dishes, linked recipes, menus with a connection to current events, etc.). Interns are primarily responsible for maintaining these accounts (under the supervision of the project directors), and this affords them the experience of acting as representatives of a collection and, at times, providers of a kind of web-based reference service.

Around the same time that the intern team was coming online, a new experimental technology and design unit was assembled at the Library: NYPL Labs. This had several implications. First, it meant shifting away from contractors to building in-house expertise. Establishing a dedicated framework for collections innovation also led the team to think more deeply about the implications of the project. Developing new user experiences would remain a strong focus, but so would data management and technical sustainability. Tackling the internal workflow issues that WOTM had exposed became higher priority—especially as public participation remained vigorous without the addition of any new website features or frills. What’s more, the project’s success sparked discussion at the highest levels of the institution about how to treat and preserve user-contributed metadata, so it became all the more incumbent upon the WOTM team to create a viable data management model and process.

Additionally, although the WOTM beta site was effective from a user standpoint, it lacked functionality that would enable users to browse, search, and explore the collection (and the data emerging out of it). Moreover, many core functions needed to be re-written in the interest of sustainability and performance. NYPL Labs decided to build Menus 2.0 for June 2012, timed to coincide with the opening of *Lunch Hour NYC*, a major exhibition held at The New York Public Library co-curated by Rebecca Federman.

We decided to focus our new development efforts on the following areas: 1) overhaul of the application code; 2) new visual design and user interface improvements; 3) new search engine and browsing features; and 4) a public API (applications programming interface) to provide other application developers and digital researchers access to real-time data from the project.

infrastructure providers such as Heroku, enabling the site to scale to meet its increased traffic without worry.



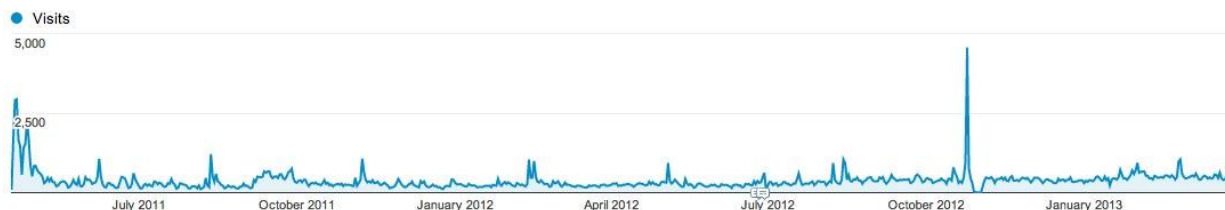
New dish page interface, with data visualization tools

Despite the success of the low-barrier approach, we sometimes lament that we were not better able to recognize the contributions of the top transcribers, or to develop a hierarchy of tasks where more challenging work might be offered to the more experienced participants. Beyond raw analytics, all of the user stories are anecdotal, but we do nevertheless have a sense that the project tends to attract people with a strong culinary or library science interest, and that judging by the traffic logs, transcription happens all throughout the day, and approximately 90% of visitors come to the site on a daily basis. We also know through web analytics that the project, while drawing predominantly from the English-speaking world (about 75% United States), has an audience that far transcends New York. A little less than a quarter of overall visits originate within the state.

Accomplishments

What's On The Menu? has won various accolades, including the prestigious Roy Rosenzweig Award for Innovation in Digital History from the American Historical Association and was widely touted as a model in the cultural heritage sphere. NYPL Labs has been awarded the Innovative Use of Archives Award from the Archivists Round Table of Metropolitan New York and given a commendation of merit from Stanford Prize for Innovation in Research Libraries on the strength of its project portfolio, including WOTM. Since the initial frenzy around launch, participation had tapered but then settled into

a steady pattern of deep engagement, with occasional attention spikes whenever significant new press or social media coverage occurred.



Web analytics provide compelling clues as to the *depth* of engagement the site has fostered. In the 23 months following launch (April 18, 2011 to March 18, 2013), page views per visitor have averaged 18.14 (compare to 2.21 for visits to NYPL’s main website in the same period). That has led 260,256 visits, producing over 4.7 million page views. Similarly, time spent on WOTM has averaged at 5:31 minutes (compared to 2:33 on NYPL.org). Among visitors who’ve transcribed a dish, results are staggering: they spend an average of 22:39 on the site and visit an average of 88.12 pages per session. This suggests a story of deep immersion in the transcription activity and in exploration of the steadily growing trove of menu-derived data.

From the earliest days of the project, we released the cumulative dataset of transcriptions to the public as a key part of the site. We always viewed WOTM as the public construction of a public resource, and consequently desired to make as much of its output as possible available to the public as soon as possible. As part of the site’s relaunch in 2012, we took several of the tools for exploring data and made them a key part of the site itself. Bi-weekly exports (containing all dishes, prices, page coordinates, and bibliographic data for each menu) are available for download. As previously noted, the Labs team also created NYPL’s first publicly promoted API (application programming interface), providing more technically-advanced users programmatic access into the data set. In the months since release, dozens of requests have come in for API access, representing initial interest from a wide range of constituencies ranging from food-focused tech startups, to computational linguistics researchers, to journalists, to library and museum hackers. We worked with the team at National Public Radio’s Planet Money to reveal what the fluctuations in the price of Filet Mignon over time say about the changes in the US economy⁹ and with staff in NYPL’s Multimedia group to show that a cup of coffee in New York is, in inflation-adjusted terms, cheaper than it’s ever been.

The future applications of the WOTM data remain to be seen. Undoubtedly, the crowdsourcing effort has raised the profile of the collection many times over, landing it frequently in the press over the past two years, and consistently generating small ripples through the culinary and techie social media subcultures. It also has radically enhanced the accessibility of the collection. A perusal of keyword-driven traffic to the site reveals a plethora of fascinatingly obscure “long-tail” searches that have brought users serendipitously to our menus:

- “eggs argenteuil” — a scrambled egg preparation found in 1910, reappearing in 1961 (88 visits);
- “remus fizz” — a citrusy gin cocktail with egg whites, mid-century (40 visits);
- “moskowitz and lupowitz” — a vanished Romanian-Jewish eatery from the Lower East Side (23 visits);
- “ss Homeric” — an ocean liner, originally German, relinquished to Britain as war reparations in 1919 (16 visits).

⁹ Lam Thuy Vo, NPR – Planet Money, “What A Very Old Menu Tells Us About The Price of Steak”, <http://www.npr.org/blogs/money/2012/08/13/158719677/a-new-york-steakhouse>

By these and other measures, we can witness the collection's weaving into the fabric of the web.

The Future: Continuation & Long-term Impact

Nearly two years into the project, it is still going strong. We're creating new features and activities, new uses for the data, and exploring new processes within the Library inspired by *WOTM*'s success.

Our most basic obligation to the project is to continue the reprocessing of the collection for digitization, so we can provide more menus to our transcribers. Our early need to digitize more menus instigated a significant change in NYPL's digitization priorities so they could tackle demand-driven needs of our patrons in rapid response, not just the long-term, large-scale efforts that formed the core of the digitization queue. Every few weeks, a fresh batch of menus appear on the site. This quest for more menus has also led to preliminary discussions with other institutions about collection federation. We envision a not-too-distant day where other libraries, archives, and museums with robust menu collections could have their collections served and transcribed through *What's On The Menu?*, enriching multiple institutions collections and building a more comprehensive and representative dataset.

The Library has also begun building new tasks to augment the data in the menus. On March 7, 2013 we launched the Menus Geotagger¹⁰, a new tool that allows our users to identify and classify the geospatial location of menus in the collection. Soon, this data will let us answer not only the question of "What were the prices of oysters over time?" but "What were the prices of oysters in New York City?" The approach we take side-steps the fact that many of the addresses of historic venues have changed or no longer exist, rendering contemporary geocoders and gazetteers inaccurate, by asking participants to identify the location to the closest level they can, be it the physical address, the city, the state, or the country, and pinpointing that location on a map zoomed to those specifications. This data will soon let the menus mingle alongside other geospatial datasets in the Library such as our historic maps, New York City street views, theatrical playbills, city and business directories etc. All of this historical geodata will be referenced via another NEH ODH-supported start-up project, *NYC Chronology of Place*, a historical gazetteer of New York City, which will enable new sorts of educational and research applications not yet imagined.

But the biggest impact the project has had has been the way it's reshaped the conception within NYPL of the role of civic participation and crowdsourcing in the way we operate. The Menus project created a level of participation with our patrons that lets them engage with the Library on their own terms and their own time. Consequently, several of the engagement patterns are being adopted into other areas of operations. This shift in thinking led to the formation of NYPL Labs, which has launched several other crowd participation applications, including Ensemble (collaborative theatre program transcription to build a dataset of historic performance), *DirectMe NYC: 1940* (a search tool for the 1940 census designed to provide users with better research workflows online while providing the library data which could be used for text corrections), and the Stereogramimator (user-created remixes of our stereographic photography collections). The philosophy has spread into many of the Library's core services, including our catalog, which now includes the ability for users to rate, tag, and contribute reviews of everything we offer, and will permeate our next-generation core collections access platform, where user participation, transcription, annotation, and augmentation will be built in as core services for most of our digital assets from the start.

¹⁰ <http://menus.nypl.org/geo>